D1018–D1022 *Nucleic Acids Research, 2015, Vol. 43, Database issue* *Published online 01 October 2014*
doi: 10.1093/nar/gku894

RiceVarMap: a comprehensive database of rice genomic variations

Hu Zhao, Wen Yao, Yidan Ouyang, Wanneng Yang, Gongwei Wang, Xingming Lian, Yongzhong Xing, Lingling Chen and Weibo Xie*

National Key Laboratory of Crop Genetic Improvement, National Center of Plant Gene Research (Wuhan), Huazhong Agricultural University, Wuhan 430070, China

Received August 11, 2014; Revised September 15, 2014; Accepted September 16, 2014

ABSTRACT

Rice Variation Map (RiceVarMap, http:/ricevarmap.ncpgr.cn) is a database of rice genomic variations. The database provides comprehensive information of 6 551 358 single nucleotide polymorphisms (SNPs) and 1 214 627 insertions/deletions (INDELs) identified from sequencing data of 1479 rice accessions. The SNP genotypes of all accessions were imputed and evaluated, resulting in an overall missing data rate of 0.42% and an estimated accuracy greater than 99%. The SNP/INDEL genotypes of all accessions are available for online query and download. Users can search SNPs/INDELs by identifiers of the SNPs/INDELs, genomic regions, gene identifiers and keywords of gene annotation. Allele frequencies within various subpopulations and the effects of the variation that may alter the protein sequence of a gene are also listed for each SNP/INDEL. The database also provides geographical details and phenotype images for various rice accessions. In particular, the database provides tools to construct haplotype networks and design PCR-primers by taking into account surrounding known genomic variations. These data and tools are highly useful for exploring genetic variations and evolution studies of rice and other species.

INTRODUCTION

Single nucleotide polymorphisms (SNPs) and insertions/deletions (INDELs) are two major types of genomic variations in living beings and have been widely utilized in basic research and industry. Researchers explored genetic variations as markers to study the genetic basis of complex traits via linkage and association mapping (1–3). High-density markers increase the probability that some markers are just located in or nearby the target genes, which would provide great advantages for marker-assisted selection in breeding (4). And in some cases, sufficient information of genetic variations may result in identifying the causal genes directly. In addition, when genotype information of individuals in a population is available, we can carry out various population genetic studies, such as estimating linkage disequilibrium blocks, constructing haplotype networks (5), inferring population history (6) and identifying natural or artificial selection signatures (7).

Rice (*Oryza sativa* L.) is both a major staple crop which feeds nearly half of the population in the world and a model species in the plant research community. Although the draft genome sequences of both *indica* and *japonica* subspecies of rice were released in 2002 (8,9) and the genome sequences of *japonica* subspecies was finished in 2005 (10), the available databases of genomic variations in rice are still very limited for bench researchers. Most of the SNP data in rice deposited in dbSNP (11) and other databases (12–14) were identified based on comparisons of the draft genome sequences between the two rice subspecies. Recently, OryzaSNP consortium interrogated 20 diverse rice varieties using resequencing microarrays and obtained 160 000 non-redundant SNPs (15). The data can be queried at a convenient database OryzaSNP (http://oryzasnp.plantbiology.msu.edu). Huang *et al.* sequenced 950 rice accessions and identified 4 109 366 non-singleton SNPs (3). They further carried out genome-wide association study (GWAS) for a lot of important agronomic traits and provided a database (http://202.127.18.221/RiceHap2/) as supporting information of the corresponding publication. The HapRice database published this year provides allele frequencies of about 3300 SNPs in 253 rice accessions (16). These newly emerging databases provide progresses in some ways. However, a comprehensive database, like HapMap (17) in human research community, is highly required in rice. Such a database should include information of abundant high-quality genomic variations and detailed genotypes of a mass of rice accessions, frequency of different subpopulations and comprehensive annotation. Meanwhile, a user-friendly query interface and useful visualization tools should be available.

*To whom correspondence should be addressed. Tel: +86 27 87281677; Fax: +86 27 87287092; Email: weibo.xie@mail.hzau.edu.cn

To achieve these objectives, we collected and analysed sequence data of 1479 rice accessions, identified 6 551 358 SNPs and 1 214 627 INDELs, and constructed a comprehensive database of rice genomic variations, RiceVarMap. Compared with extant relevant databases, RiceVarMap provides not only the largest set of genomic variations and genotype data at present, but also characteristic phenotype images for various rice accessions and a set of intuitional query interface and useful tools, such as constructing haplotype networks and designing polymerase chain reaction (PCR) primers by taking into account surrounding known genomic variations. These data and tools are highly useful for exploring genetic variations and evolution studies of rice and other species. Based on these data, we also investigated natural variations in rice metabolism using GWAS approach (18) and designed two genotyping arrays for rice breeding (19,20). These data will be added to the database in the future.

DATA COLLECTION, PROCESSING AND EVALUATION

Currently, we collected sequencing data from two sets of rice germplasms comprising a total of 1479 accessions of cultivated rice (*Oryza sativa* L.). The first set of germplasm consists of 529 accessions selected to represent both the usefulness in rice improvement and the genetic diversity in the cultivated species (21,22). We sequenced the 529 accessions using the Illumina HiSeq 2000 in the form of 90-bp paired–end reads to generate high-quality sequences of more than one gigabase per accession (>2.5× per genome, total 6.7 billion reads). These raw data are available in NCBI with BioProject accession number PRJNA171289. The second set of germplasm is 950 rice accessions sequenced by Huang *et al.* (3) that were downloaded from the EBI European Nucleotide Archive (accession number ERP000106 and ERP000729), which consists of 4.6 billion 73-bp paired-end reads (∼1× per genome). Together these two sets of germplasms include both landraces and improved varieties from 73 countries. The two sets of sequences provide ∼2400-fold coverage of the rice genome.

The detailed procedures of data analysis could be found in our published article (18) and webpages of RiceVarMap. Briefly, the two sets of raw sequences were combined together and aligned to rice reference genome (Nipponbare, MSU version 6.1) (23) using Burrows-Wheeler Aligner (BWA) (24). A total of 6 551 358 high-quality SNPs and 1 214 627 INDELs were identified using SAMtools and BCFtools (25). After obtaining raw genotype calls from BCFtools, 47.1% of genotypes were missing due to low-coverage sequencing. We then performed imputation using an in-house modified *k* nearest neighbour algorithm (26,27), resulting in an overall missing data rate reduced to 0.42% after imputation. A total of 6 428 770 SNPs with genotype missing data rate less than 20% were used for GWAS (with an overall missing data rate of 0.38%) (18). To estimate the accuracy of imputed genotypes, we genotyped 48 accessions using Illumina Infinium array RiceSNP50 (19). The results suggested an accuracy of 99.3% (the details of evaluation can be found in the website). Thus, the missing data rates and accuracy of the imputed genotype data set were comparable to high-coverage sequencing results.

We further inferred the population structure of the 1479 accessions using ADMIXTURE (28). The accessions were accordingly classified into 809 *indica*, 547 *japonica*, 67 *Aus* and 56 intermediate type. The *indica* could be further classified into *indica I* which has germplasms of South China origin, *indica II* which contains germplasms from Southeast Asia and *indica* intermediate type. The *japonica* could also be divided into temperate *japonica*, tropical *japonica* and *japonica* intermediate type. The population structure of the collected rice accessions is shown in Figure 1a. The allele frequencies of each SNP in different populations by different classification were calculated and stored in RiceVarMap.

DATABASE FEATURES

RiceVarMap is free and open to the public with comprehensive functions (Figure 1b). More detailed information is described as follows. In the database, each SNP or INDEL is labelled with a unique identifier (ID, e.g. sf0100000131, vf0136465397). The first letter of the ID indicates the polymorphic type, 's' for SNP and 'v' for INDEL. The second letter represents the version of the reference genome, 'f' for the version 6.1 of Nipponbare. The number is the chromosome coordinate of a variation, e.g. sf0100000131 means a SNP at chromosome 1, 131 bp.

Search for SNPs/INDELs by region

Information of SNPs/INDELs can be queried by limiting genomic coordinates of the rice genome. Since the reference genome used by RiceVarMap is the version 6.1 of Nipponbare from MSU, it should be ensured that all coordinates correspond to this version before query. Basic Local Alignment Search Tool search in the 'Tools' menu can be used to obtain corrected genomic coordinates from sequence. Furthermore, SNPs can be filtered by limiting ranges of allele frequencies in different populations (maximum three simultaneous combinations).

Search for SNPs/INDELs within gene

The SNPs/INDELs may have great influence on gene functions. We provide a function for users to search SNPs/INDELs by gene identifiers, gene symbols or keywords of gene annotation, and wildcard characters are accepted. In order to better explain the functional changes caused by SNPs/INDELs, we utilized SNP effector (29) to annotate SNPs/INDELs and the ones with large-effect changes would be highlighted in the result page. Moreover, user can define upstream and downstream regions of genes and retrieve SNPs/INDELs in these regions. The acquired SNPs/INDELs would be displayed in the result page with the structure of the gene in a graph.

Search for genotypes with SNP/INDEL ID

In this interface, users can fetch the genotypes of different accessions through entering SNP/INDEL ID and selecting

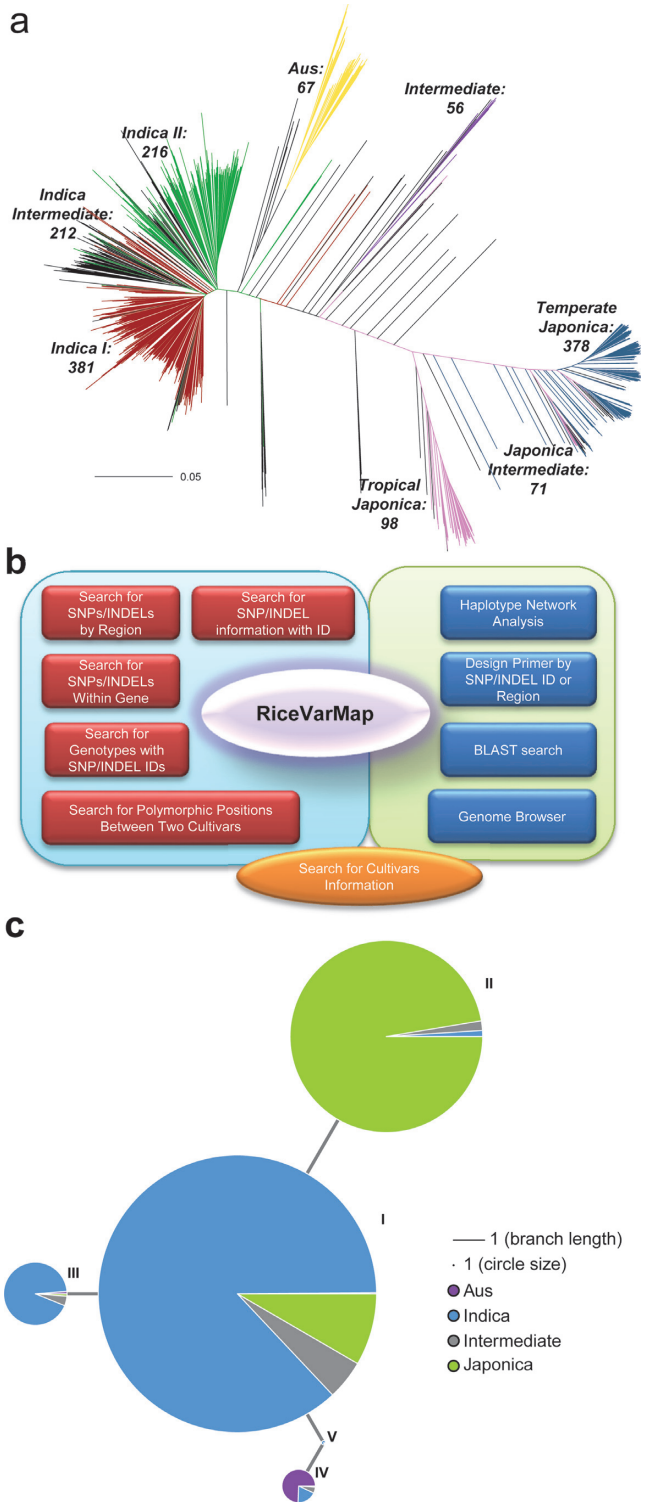

Figure 1. Contents and functions of RiceVarMap. (**a**) Population structure of the 1479 rice accessions included in RiceVarMap. Labels denote the names of subpopulations and the number of accessions in the subpopulation. (**b**) The functions of RiceVarMap. (**c**) Haplotype network of *S5* constructed using the tool 'Haplotype Network Analysis' implemented in RiceVarMap. Each circle represents a haplotype and the size is proportionate to the number of accessions with that haplotype. Branch length represents the genetic distance between two haplotypes.

corresponding accessions. If the target genotype is a minor allele, it will be displayed in red in the search results. The genotypes can be downloaded in csv format.

Search for SNP/INDEL information with ID

Detailed information of single SNP/INDEL is described in this page, which includes not only basic SNP/INDEL information (e.g. position, major allele, minor allele) but also allele frequencies in different populations and information of SNP/INDEL variation effects.

Search for polymorphic positions between two cultivars

It is easy to figure out the polymorphic SNPs between two cultivars in this page, which makes it convenient to generate new molecular markers for bench work in further study.

Search for information of the cultivars

The geographical information of 1479 rice accessions is provided, and each cultivar is assigned with a cultivar ID (e.g. C001, HP84 and W001). The classification information and phenotype images of these rice accessions are available from this page. At the same time, we use Google map to help users to locate the position of each accession.

Design primer by SNP/INDEL ID

RiceVarMap also provides tools to facilitate bench works and further analysis. This tool is provided for researchers to pick PCR primers to validate SNPs/INDELs or develop molecular markers. Primers will be designed to flank the target SNP/INDEL and avoid to overlap with known SNPs/INDELs.

Design primer by region

This tool is designed for researchers to pick PCR primers to amplify target genomic regions and avoid to overlap with known SNPs/INDELs.

Haplotype network analysis

Haplotype network is frequently used in population genetic analysis. RiceVarMap provides a simple tool to generate haplotype network using modified functions from R package pegas (30). Users can download PDF format graph and the detailed haplotype information for further analysis.

A CASE OF APPLICATION

Hybrid sterility is a common phenomenon between different populations; one of the best-known examples for hybrid sterility is the case between the two rice subspecies, *indica* and *japonica*. The *S5* locus has been well characterized to regulate fertility in *indica-japonica* hybrids, which are encoded by three tightly linked genes (31,32). We used one of the genes, LOC_Os06g11010, a 'killer' of gametes, to demonstrate the functions of RiceVarMap. A total of 14 SNPs were found from 'Search for SNPs within gene'

function using the gene locus name as input. Five of them (sf0605759642, sf0605759919, sf0605760007, sf0605760352, sf0605760512) with minor allele frequencies greater than 0.05 in the whole population could be identified easily from the results and we observed that all of them were non-synonymous mutations, suggesting that this gene might undergo rapid evolution. We then copied IDs of these SNPs and generated a haplotype network using the tool 'Haplotype Network Analysis' (Figure 1c). There are five haplotypes found with at least 10 accessions and we found that nearly all *aus* accessions were of haplotype IV, which was characterized as wide-compatibility accessions that could overcome the sterility of the hybrid. Thus, accessions with this haplotype would be very useful for hybrid breeding. Finally, we could use the tool 'Design Primer by SNP/INDEL ID' to develop molecular markers based on these SNPs as well.

FUTURE DEVELOPMENT

We will make efforts to improve and update the database from the following three aspects. First, as more rice accessions are sequenced (6,33) and publicly available, the size of the database will be enlarged with more rice accessions, SNPs/INDELs and genotypes. And the reference genome used will be updated as well. Second, we have identified thousands of significant loci regulating metabolism (18) and constructed high-throughput phenotyping platforms (34). We are planning to add these data into RiceVarMap, making it a comprehensive database of rice genomic, metabolomic and phenomic variations. Third, we will take efforts to make the database more user-friendly and more efficient with the reflection and feedback of the first version of RiceVarMap.

ACKNOWLEDGEMENTS

We thank Prof. Sibin Yu and Mr. Zilong Guo for helping to prepare plant photos, and Mr. Zuoxiong Liu for English editing. We also thank anonymous reviewers for the help in improving our database and the manuscript.

FUNDING

National High Technology Research and Development Program of China (863 Program) [2012AA10A304, 2014AA10A602]; National Natural Science Foundation of China [31100962, 31123009, J1103510]; Fundamental Research Funds for the Central Universities [2011PY068]. Funding for open access charge: National High Technology Research and Development Program (863 Program) [2012AA10A304], the Ministry of Science and Technology of China.

Conflict of interest statement. None declared.

REFERENCES

1. Buckler,E.S., Holland,J.B., Bradbury,P.J., Acharya,C.B., Brown,P.J., Browne,C., Ersoz,E., Flint-Garcia,S., Garcia,A., Glaubitz,J.C. *et al.* (2009) The genetic architecture of maize flowering time. *Science*, **325**, 714–718.
2. Zhao,K., Tung,C.W., Eizenga,G.C., Wright,M.H., Ali,M.L., Price,A.H., Norton,G.J., Islam,M.R., Reynolds,A., Mezey,J. *et al.* (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat. Commun.*, **2**, 467.
3. Huang,X., Zhao,Y., Wei,X., Li,C., Wang,A., Zhao,Q., Li,W., Guo,Y., Deng,L. and Zhu,C. (2012) Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat. Genet.*, **44**, 32–39.
4. Xu,Y. and Crouch,J.H. (2008) Marker-assisted selection in plant breeding: from publications to practice. *Crop Sci.*, **48**, 391–407.
5. Lu,L., Yan,W., Xue,W., Shao,D. and Xing,Y. (2012) Evolution and association analysis of Ghd7 in rice. *PLoS ONE*, **7**, e34021.
6. Huang,X., Kurata,N., Wei,X., Wang,Z.X., Wang,A., Zhao,Q., Zhao,Y., Liu,K., Lu,H., Li,W. *et al.* (2012) A map of rice genome variation reveals the origin of cultivated rice. *Nature*, **490**, 497–501.
7. Hufford,M.B., Xu,X., van Heerwaarden,J., Pyhajarvi,T., Chia,J.M., Cartwright,R.A., Elshire,R.J., Glaubitz,J.C., Guill,K.E., Kaeppler,S.M. *et al.* (2012) Comparative population genomics of maize domestication and improvement. *Nat. Genet.*, **44**, 808–811.
8. Yu,J., Hu,S., Wang,J., Wong,G.K., Li,S., Liu,B., Deng,Y., Dai,L., Zhou,Y., Zhang,X. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science*, **296**, 79–92.
9. Goff,S.A., Ricke,D., Lan,T.H., Presting,G., Wang,R., Dunn,M., Glazebrook,J., Sessions,A., Oeller,P., Varma,H. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science*, **296**, 92–100.
10. International Rice Genome Sequencing Project. (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.
11. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
12. Feltus,F.A., Wan,J., Schulze,S.R., Estill,J.C., Jiang,N. and Paterson,A.H. (2004) An SNP resource for rice genetics and breeding based on subspecies *indica* and *japonica* genome alignments. *Genome Res.*, **14**, 1812–1819.
13. Shen,Y.J., Jiang,H., Jin,J.P., Zhang,Z.B., Xi,B., He,Y.Y., Wang,G., Wang,C., Qian,L., Li,X. *et al.* (2004) Development of genome-wide DNA polymorphism database for map-based cloning of rice genes. *Plant Physiol.*, **135**, 1198–1205.
14. Zhang,Z., Deng,Y., Tan,J., Hu,S., Yu,J. and Xue,Q. (2007) A genome-wide microsatellite polymorphism database for the *indica* and *japonica* rice. *DNA Res.*, **14**, 37–45.
15. McNally,K.L., Childs,K.L., Bohnert,R., Davidson,R.M., Zhao,K., Ulat,V.J., Zeller,G., Clark,R.M., Hoen,D.R. and Bureau,T.E. (2009) Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 12273–12278.
16. Yonemaru,J., Ebana,K. and Yano,M. (2014) HapRice, an SNP haplotype database and a web tool for rice. *Plant Cell Physiol.*, **55**, e9.
17. The International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.
18. Chen,W., Gao,Y., Xie,W., Gong,L., Lu,K., Wang,W., Li,Y., Liu,X., Zhang,H., Dong,H. *et al.* (2014) Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nat. Genet.*, **46**, 714–721.
19. Chen,H., Xie,W., He,H., Yu,H., Chen,W., Li,J., Yu,R., Yao,Y., Zhang,W., He,Y. *et al.* (2014) A high-density SNP genotyping array for rice biology and molecular breeding. *Mol. Plant*, **7**, 541–553.
20. Yu,H., Xie,W., Li,J., Zhou,F. and Zhang,Q. (2014) A whole-genome SNP array (RICE6K) for genomic breeding in rice. *Plant Biotechnol. J.*, **12**, 28–37.
21. Yu,S.B., Xu,W.J., Vijayakumar,C.H., Ali,J., Fu,B.Y., Xu,J.L., Jiang,Y.Z., Marghirang,R., Domingo,J., Aquino,C. *et al.* (2003) Molecular diversity and multilocus organization of the parental lines used in the International Rice Molecular Breeding Program. *Theor. Appl. Genet.*, **108**, 131–140.
22. Agrama,H., Yan,W., Lee,F., Fjellstrom,R., Chen,M.-H., Jia,M. and McClung,A. (2009) Genetic assessment of a mini-core subset developed from the USDA rice genebank. *Crop Sci.*, **49**, 1336–1346.
23. Ouyang,S., Zhu,W., Hamilton,J., Lin,H., Campbell,M., Childs,K., Thibaud-Nissen,F., Malek,R.L., Lee,Y., Zheng,L. *et al.* (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res.*, **35**, D883–D887.

24. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
25. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
26. Roberts,A., McMillan,L., Wang,W., Parker,J., Rusyn,I. and Threadgill,D. (2007) Inferring missing genotypes in large SNP panels using fast nearest-neighbor searches over sliding windows. *Bioinformatics*, **23**, i401–i407.
27. Huang,X., Wei,X., Sang,T., Zhao,Q., Feng,Q., Zhao,Y., Li,C., Zhu,C., Lu,T., Zhang,Z. *et al.* (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.*, **42**, 961–967.
28. Alexander,D.H., Novembre,J. and Lange,K. (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.*, **19**, 1655–1664.
29. Cingolani,P., Platts,A., Wang,L.L., Coon,M., Nguyen,T., Wang,L., Land,S.J., Lu,X. and Ruden,D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; *iso-2*; *iso-3*. *Fly*, **6**, 80–92.
30. Paradis,E. (2010) pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics*, **26**, 419–420.
31. Chen,J., Ding,J., Ouyang,Y., Du,H., Yang,J., Cheng,K., Zhao,J., Qiu,S., Zhang,X. and Yao,J. (2008) A triallelic system of S5 is a major regulator of the reproductive barrier and compatibility of indica–japonica hybrids in rice. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 11436–11441.
32. Yang,J., Zhao,X., Cheng,K., Du,H., Ouyang,Y., Chen,J., Qiu,S., Huang,J., Jiang,Y. and Jiang,L. (2012) A killer-protector system regulates both hybrid sterility and segregation distortion in rice. *Science*, **337**, 1336–1340.
33. The 3000 Rice Genomes Project. (2014) The 3,000 rice genomes project. *Gigascience*, **3**, 7.
34. Yang,W., Duan,L., Chen,G., Xiong,L. and Liu,Q. (2013) Plant phenomics and high-throughput phenotyping: accelerating rice functional genomics using multidisciplinary technologies. *Curr. Opin. Plant Biol.*, **16**, 180–187.